DS-GA 3001: Introduction to causal inference for data scientists

Course description: Causal inference is the science of analyzing causal relationships between events. What is the impact of advertising on demand? By how much will a graduate degree increase one’s salary? What is the impact of minimum wage on wages and unemployment? These are questions that require the understanding of causal connections between a decision and its consequences. Surprisingly enough, causal inference tools have only been relatively recently developed and taken to data. We will present Rubin’s model of potential outcomes, which will be our primary framework in the course. We will study cases when the analyst has the power to design a randomized experiment (Randomized control trials, field experiments, AB testing) and cases when the analyst is present with a dataset where randomization has not explicitly occurred (observational studies). Our course will close with an overview of special topics in causal inference, such as causal inference in networks, and successful examples of machine learning for causal inference.

Instructor: Michael Gill (NYU CDS), mzgill@nyu.edu, Office hours: Tuesdays, 3pm-4:30pm, CDS, Room 620
Teaching assistant: Lei Xu (NYU CDS), lx557@nyu.edu, Office hours: Wednesdays, 1:30pm-3pm, CDS, Room 665

Schedule:
Lecture: Tuesdays, 1pm-2:40pm, 60 Fifth Avenue, Room C12
Lab: Wednesdays, 8:35pm-9:25pm, 60 Fifth Avenue, Room C10

Prerequisites: A first course in probability and statistics, and be acquainted with linear algebra.

Evaluation: Class participation/attendance (10%), Homework (35%), Midterm (25%), Final Exam (30%).

Textbooks:
There is NO required textbook. However, recommended books include:

Software: The core of the course will be taught using the statistical computing environment R, freely available from https://www.r-project.org/. If students prefer to use Python instead, that is fine, but students will be expected to submit working Jupyter notebooks on problem sets. In the first lab session we will help students install R Studio.

Datasets: will be available on the course webpage. Datasets used in lecture and lab include, but are not limited to:
- Lalonde training program evaluation data
- Bertrand-Mullainathan labor market discrimination data
- Angrist data on draft Vietnam lottery

Homework: There will be 5 problem sets throughout the semester, released every two weeks. They are not meant to be long or onerous, but to help clarify concepts relayed in lecture and lab. Students will be expected to submit problem sets as PDFs, and (when appropriate) share replication code alongside the writeup. Students that fail to submit both write-ups and replication code will not receive full marks on their homework.

Homework policies:
• Late assignments will receive a 10% deduction for each hour late submitted after the posted deadline, rounding up. For example, an assignment submitted 45 minutes late will receive a 10% deduction, while an assignment submitted 70 minutes late will receive a 20% deduction. However, there will be a 10 minute “grace period” before an assignment is considered late. Requests for homework extensions outside of these terms will not be granted. Please contact the teaching staff for any additional questions relating to this policy.
• Student collaboration on problem sets is not permitted. It is okay to ask each other clarifying questions about concepts relevant to the homework, but students should never explicitly share their work or answers with others.

Exams: There will be two exams throughout the semester: a midterm on 3/6/2018, and a final on 5/15/2018. Both will be closed notes, and no calculators allowed. Please note the timing of the final exam may change, but the midterm time is fixed. The final exam date/time is tentative estimate from the NYU Registrar (as of Jan. 17, 2018).
Part I: Introduction

LECTURE 1. CAUSAL INFERENCE: SOME MOTIVATING EXAMPLES. 1/23/2018
• MW, Chapter 1

LECTURE 2. THE NEYMAN-RUBIN MODEL OF POTENTIAL OUTCOMES. 1/30/2018
• MW, Chapter 2.
• IR, Chapter 1.
• HR, Chapter 1.
  ◦ Homework 1 announced in lab, due by 8:30 pm on 2/7/2018.

Part II. Causal inference in an experimental setting

LECTURE 3. RCTs, AB TESTING, BUSINESS EXPERIMENTS (1). 2/6/2018
• IR, Chapter 4.
  ◦ Homework 1 due by 8:30 pm on 2/7/2018.

LECTURE 4. RCTs, AB TESTING, BUSINESS EXPERIMENTS (2). 2/13/2018

LECTURE 5. NONCOMPLIANCE & INSTRUMENTAL VARIABLES (1). 2/20/2018
• MW, Chapter 9
• IR, Chapters 23, 24, 25.
  ◦ Homework 2 announced in lab, due by 8:30 pm on 2/28/2018.

Part III. Causal inference in observational studies

LECTURE 6. INSTRUMENTAL VARIABLES (2) & INTRO TO OBSERVATIONAL STUDIES. 2/27/2018
• MW, Chapter 5
• IR, Chapters 12-13.
  ◦ Homework 2 due by 8:30 pm on 2/28/2018.

*Midterm Exam: Tuesday, March 6, 2018. In class.*

LECTURE 7. MATCHING ESTIMATORS. 3/20/2018
LECTURE 8. DIFFERENCES-IN-DIFFERENCES, REGRESSION DISCONTINUITY.

• MW, Chapters 6-7
  o Homework 3 announced in lab, due by 8:30 pm on 4/4/2018.

LECTURE 9. EXTENDING DIFFERENCES-IN-DIFFERENCES.

  o Homework 3 due by 8:30 pm on 4/4/2018.

LECTURE 10. HIGH-DIMENSIONAL MODELS.


Part IV. Special topics

LECTURE 11. PRACTICAL CHALLENGES WITH INFERENCE.

  o Homework 4 announced in lab, due by 8:30 pm on 4/25/2018.

LECTURE 12. CAUSAL INference IN NETWORKS.

  o Homework 4 due by 8:30 pm on 4/25/2018.

LECTURE 13. MACHINE LEARNING AND CAUSAL INFERENCE.

  o Homework 5 announced in lab, due by 8:30 pm on 5/8/2018.

*Reading Week: No class, but Homework 5 due by 8:30 pm on 5/8/2018*

*Final Exam: Tuesday, May 15 from 2-3:50pm. Date and timing subject to change, from NYU Registrar*